COMBINING HARMONY-BASED AND NOVELTY-BASED APPROACHES FOR STRUCTURAL SEGMENTATION

Johan Pauwels, Florian Kaiser and Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
johan.pauwels@ircam.fr, florian.kaiser@ircam.fr, geoffroy.peeters@ircam.fr

ABSTRACT

This paper describes a novel way to combine a well-proven method of structural segmentation through novelty detection with a recently introduced method based on harmonic analysis. The former system works by looking for peaks in novelty curves derived from self-similarity matrices. The latter relies on the detection of key changes and on the differences in prior probability of chord transitions according to their position in a structural segment. Both approaches are integrated into a probabilistic system that jointly estimates keys, chords and structural boundaries. The novelty curves are herein used as observations. In addition, chroma profiles are used as features for the harmony analysis. These observations are then subjected to a constrained transition model that is musically motivated. An information theoretic justification of this model is also given. Finally, an evaluation of the resulting system is performed. It is shown that the combined system improves the results of both constituting components in isolation.

1. INTRODUCTION

Structural segmentation of music is the process in which an audio recording is divided into a number of non-overlapping sections that correspond to the macro-temporal organisation of a piece. These entities usually take the form of verses and choruses in popular music, or of movements in classical music. The obtained sections can then be used for interactive listening, audio summarization, synchronization or as an intermediate step in further content-based indexing.

Traditional approaches to structural segmentation have been categorized into three categories [14]: repetition-based, novelty-based and homogeneity-based methods. A mid-level representation, called self-similarity matrix, is often used for these task. It is obtained from the feature sequence by comparing each instance with all time-delayed copies of itself according to some similarity measure. The result is a visualisation of the musical structure. It was originally introduced into the music domain by Foote [2]. Repetition-based approaches rely on the hypothesis that recurring patterns in the feature sequence cause a perception of a higher structure. In a self-similarity matrix this becomes visible as stripes on the off-diagonals [4]. Novelty-based systems try to identify transitions between two contrasting parts, which are also perceived as structural boundaries by humans [1]. Initially, these methods were the dual approach of homogeneity-based methods [10], as one was looking exclusively for transitions between two distinct sections that are similar according to some musical property [3]. Recently however, this approach has been extended to also include contrast between a homogeneous and a non-homogeneous section [6].

The method we propose builds upon the novelty-based method of [3], but integrates this with a novel approach that is based on the estimated harmony of the piece [16]. Previous efforts of deriving a structural segmentation from harmony have mostly been concerned with using chroma features for the construction of a self-similarity matrix, instead of or in addition to timbre-related features [5]. We however work with a higher-level harmony description, in the form of key and chord estimates. In this sense, our approach is somewhat similar to previous systems by Maddage [11] or Lee [8], but in contrast to their systems, ours works simultaneously, not sequentially. They first extract key and chord estimates, which they subsequently use as inputs for a structure estimation. A sequential system can also be constructed the other way around, using an estimate of the structure as input to aid with chord estimation. An example of this kind is the method of Mauch [12]. We, on the other hand, construct a probabilistic system that jointly estimates keys, chords and structural boundaries. It is based on the assumption that some chord combinations are more common around structural boundaries. This is especially clear when they are expressed as relative chords in a key, as this gives a musicologically richer representation. These relative chord combinations will then be used as evidence for structural boundaries, together with the positions of key changes and peaks in the novelty measure.

In the remainder of this paper, we'll first give an outline of our probabilistic system in Section 2.1. Then we'll go deeper into the details of how to integrate the harmony-based and the novelty-based approach into this framework in Section 2.2, respectively Section 2.3. Afterwards, we describe the experiments we performed and analyse the results in Section 3. We conclude with some closing remarks and ideas for future work in Section 4.

2. A PROBABILISTIC SYSTEM FOR THE JOINT ESTIMATION OF KEYS, CHORDS AND STRUCTURAL BOUNDARIES

2.1 Overview

In this section we will describe the probabilistic system that we propose for the determination of a structural segmentation of a track, along with an estimation of the local keys and chords. As a starting point we use the system of Pauwels et al. [15] for the simultaneous estimation of keys and chords. It consists of an HMM in which each state represents a combination of a key and a chord. We extend it by letting each state q represent a structural position in addition to a key and a chord. A key k can take one of N_k values, a chord one of N_c values and the structural positions s can take one of two values: L which means that q is the last state of a structural segment or O which means that it is not. Finally, we add a single state to handle the case when no chord is being played, notably at the beginning and end of a recording. In this state, the key will accordingly take a "no-key" value and the structural position will take a value of $s = R$. In summary, $q = (s, k, c)$ with $s \in \{L, O\}, k \in \{K_1, \ldots, K_{N_k}\}, c \in \{C_1, \ldots, C_{N_c}\}$ or $q = (R, \text{no-key}, \text{no-chord})$.

Finding the most likely sequence of key, chord and structure labels given a sequence of observations $X = \{x_1, x_2, \ldots, x_T\}$ then amounts to finding the state sequence $\hat{Q} = \{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T\}$ that optimally explains these observations. Because the state variable q consists per definition of the combination of a chord, key and structure variable, these three optimal sequences will always be jointly estimated. Afterwards, the structural boundaries $\hat{S}$ can be derived from the optimal state sequence by inserting a boundary for every transition from a state where $s = L$ to one where $s = O$, or from or to a state with $s = R$. The derivation of the optimal key $\hat{K}$ and chord sequence $\hat{C}$ from the latter is even more trivial.

By applying Bayes' theorem, and further assuming the first order Markov property and independence of the observations, we can rewrite the probability to be maximized to

$$\hat{S}, \hat{K}, \hat{C} = \arg\max \prod_{t=1}^{T} P(\mathbf{x_t}|s_t, k_t, c_t) \\ P(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1})$$

The time-interval t here indicates the index of the interbeat segments, where the beats are estimated by ircambeat [17]. The probabilities of this HMM will now be determined by the combination of 2 different components that are each using their separate observations. The first component is a method to estimate a structural segmentation simultaneously with the harmony of a piece. It uses chroma features. The second component determines structural boundaries based on a novelty measure that is derived from timbral features. For clarity reasons, we introduce separate notations for the chroma observations $\mathbf{y_t}$ and for the novelty measure $\mathbf{z_t}$ (so $\mathbf{x_t} = [\mathbf{y_t}\mathbf{z_t}]$). We will consider the novelty observations independent of the chroma observations: $P(\mathbf{x_t}|s_t, k_t, c_t) = P(\mathbf{y_t}|s_t, k_t, c_t)\, P(\mathbf{z_t}|s_t, k_t, c_t)$. The final probability to be maximized will therefore be

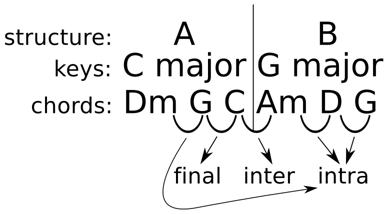

Figure 1. An example annotated sequence with the three structure dependent positions indicated

$$\hat{S}, \hat{K}, \hat{C} = \arg\max \prod_{t=1}^{T} P(\mathbf{y_t}|s_t, k_t, c_t)\, P(\mathbf{z_t}|s_t, k_t, c_t) \\ P(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1})$$

Of these three terms, the chroma observation probability $P(\mathbf{y_t}|s_t, k_t, c_t)$ and the transition probability $P(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1})$ will be set by the harmony-based component of our system, while the novelty observation probability $P(\mathbf{z_t}|s_t, k_t, c_t)$ will be set by the novelty-based component.

We use different observations for both components to capture different types of information, in hopes of them being complementary. For instance, a change of instrumentation won't be detected as a structure boundary by the harmony-based component, while the novelty-based component won't recognize a chord sequence that is typical for an ending. The choice of their respective observations reflects this.

2.2 The harmony-based component

2.2.1 Motivation and information theoretical justification

The basic premise upon which our approach is built, is that chord sequences, and more specifically chord pairs, exhibit a different prior probability depending on their position with respect to structural boundaries. In addition to the specificity of these chord combinations, we also argue that the number of distinct chord pairs at the end of a segment is lower compared to all possible chord sequences in the middle of a structural segment. Structural boundaries seem an implausible place to experiment with some less common chord combinations, established by well-known, diatonic chord combinations. Real world examples that support these statements are the observation that movements in classical music typically end with one of a select number of chord combinations, called cadences, or that musical section changes in jazz and blues are often preceded by so-called "turn-arounds".

In order to verify these statements in a methodological way, we first identify three categories of chord pairs based on their position with respect to structural boundaries. An example sequence for which these three categories are indicated can be found in Figure 1. The first class is named *final* and this contains the chord pairs that form the two last chords of a structural segment. The second category we'll

	Isophonics		Quaero	
	major	minor	major	minor
intra	6.17	6.25	4.29	4.98
inter	3.91	4.52	2.99	2.37
final	2.91	3.51	2.36	3.18

Table 1. Perplexity of relative chord transition models per mode according to structural position

call *inter* and this consists of all chord pairs that straddle a structural boundary. The last class of chord pairs is called *intra* and this includes all the remaining chord pairs, the ones that occur in the beginning and middle of a structural segment.

To reason about chord pairs in a musicologically more informative way, we will interpret them as relative chords interpreted in a key. This representation reflects more closely the way scholars analyse harmonic movement. Also in accordance to musicological analysis, both chords will be interpreted in the same key. Because of the forward motion in music, this will be the key annotated at the time of the first chord. In mathematical notation, we will define a key k as the combination of a tonic t and a mode m. A chord c is defined as the combination of a root r and a type p. Both t and r belong to one of the 12 different pitch classes. We restrict ourselves in this paper to 2 modes and 4 chord types ($m \in \{\mathrm{major}, \mathrm{(natural)minor}\}$, $p \in \{\mathrm{maj}, \mathrm{min}, \mathrm{dim}, \mathrm{aug}\}$). We then define a relative chord c' with respect to a key k by expressing the root r as the interval between the tonic and the root: $i = d(t, r)$. Therefore we can equivalently express a key-chord pair as a key-relative chord pair $(k, c) = (t, m, r, p) = (t, m, i, p) = (k, c')$. In order to take the inherent shift-invariance in harmony analysis into account, we just ignore the tonic of the key and only keep the mode. For sequences of 2 key-chord pairs, we end up with just a mode and a pair of relative chords as a representation for the local harmony: $\left(m_n, c'_n, c'_{n+1}\right)$ where n is the chord index.

We then construct relative chord transition models for each of the three structural categories by counting occurrences of all successions of 2 consecutive relative chords in a corpus that is annotated with keys, chords and structural segments. Sequences that do not appear in the data set are assigned a probability using Kneser-Ney smoothing [7]. The corpus should be annotated such that all the positions are indicated where at least one of key, chord or structural segment changes. We have two such data sets at our disposal. The first one is the publicly available "Isophonics" set and more specifically the subset that has been used for the MIREX 2010 chord estimation competition. It consists of 217 full songs, mostly by the Beatles (180 songs), the remainder by Queen (20) and Zweieck (17). The second one is a private data set called "Quaero" and it contains 53 songs from a number of diverse artists in the popular genre. We can now quantify the difference in relative chord distribution between the various structure dependent transition models by calculating the bigram model perplexity $PP\left(C'_1, C'_2|m\right)$ per mode m for each of them. C'_1 and C'_2 represent the collection of all relative chords that appear as first, respectively second, element in the bigrams. The model perplexity is defined as the exponential of the entropy $H\left(C'_1, C'_2|m\right)$ expressed in nats:

$$\begin{aligned} PP(C'_1, C'_2|m) &= \exp\left(H\left(C'_1, C'_2|m\right)\right) \\ &= \exp\left(-\sum_{c'_1, c'_2} P\left(c'_1, c'_2|m\right) \log P\left(c'_2|c'_1, m\right)\right) \\ &= \exp\Bigg(-\sum_{c'_1} P\left(c'_1|m\right) \\ &\qquad \sum_{c'_2} P\left(c'_2|c'_1, m\right) \log P\left(c'_2|c'_1, m\right)\Bigg) \end{aligned}$$

This expresses the mean prior uncertainty of a bigram as a function of its mode. A lower value means that the transition probability is concentrated into fewer combinations of two chords. The perplexities for both data sets can be found in Table 1 and as can be seen, they confirm our hypothesis: the values for the "intra"-model are indeed significantly higher than those for the "inter" and "final"-model and this for both corpora.

For the calculation of the transition probabilities in the following paragraphs, we will make use of the information captured in these structure dependent relative chord transition models. The reasoning will therefore be reversed: instead of showing that a structural boundary often suggests a specific set of relative chord pairs, we'll use the occurrence of such relative chord combinations as evidence to estimate structural boundaries.

2.2.2 *Transition probabilities*

The transition probabilities $P(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1})$ are calculated by a prior musicological model that consists of a number of submodels. By introducing some musicologically motivated constraints to the transition probabilities, we want to enforce a number of relationships between the concepts of keys, chords and structural segments. These will ensure that our estimation always produces sensible results and have as an added benefit that this also speeds up the calculation. The first three constraints we impose are 1) a key change $k_t \neq k_{t-1}$ is only allowed to occur together with a chord change $c_t \neq c_{t-1}$, 2) a structural segment must contain at least two different chords (or a single no-chord), 3) there must be a change in chord or in key between segments. These three limitations can be easily enforced by ensuring that every state change implies a chord change. This makes the state duration model effectively a chord duration model that we control by a single parameter P_s:

$$P\left(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1}\right) = \begin{cases} P_s & s_t = s_{t-1} \wedge k_t = k_{t-1} \\ 0 & s_t \neq s_{t-1} \vee k_t \neq k_{t-1} \end{cases}, \forall c_t = c_{t-1} \quad (1)$$

We use the same value for P_s as found in the original system [15].

The remaining probabilities $P(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1}), \forall c_t \neq c_{t-1}$ of the chord changing transitions are calculated by the combination of three submodels. We further apply Bayes' theorem repeatedly to arrive at a decomposition into three terms

$$\begin{aligned} &P\left(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1}\right) \\ &\quad = P\left(s_t|s_{t-1}, k_{t-1}, c_{t-1}\right) P\left(c_t|s_t, s_{t-1}, k_{t-1}, c_{t-1}\right) \\ &\qquad P\left(k_t|s_t, s_{t-1}, k_{t-1}, c_t, c_{t-1}\right) \end{aligned}$$

The first term $P\left(s_t|s_{t-1}, k_{t-1}, c_{t-1}\right)$ will be used to control the ease of changing the structure variable s and thus to control the insertion rate of segment boundaries. We use a simple model that ignores the key and chord influence and consists of a single parameter ω that balances the probability of going to $s = O$ or $s = L$ after leaving $s = O$.

$$P\left(s_t|s_{t-1}\right) = \begin{cases} \omega & s_{t-1} = O, s_t = O \\ 1 - \omega & s_{t-1} = O, s_t = L \\ 1 & s_{t-1} = L, s_t = O \\ 0 & s_{t-1} = L, s_t = L \end{cases}$$

We can already recognize the structure-dependent relative chord transition model of the previous section in the second term $P\left(c_t|s_t, s_{t-1}, k_{t-1}, c_{t-1}\right)$. Our three categories of chord transitions – *inter*, *intra* and *final* – each correspond to a certain combination of the state variables. The *intra* model will be used when $s_{t-1} = O$ and $s_t = O$, *inter* when $s_{t-1} = L$ and $s_t = O$ and *final* when $s_{t-1} = O$ and $s_t = L$. Finally, from our definition of L it follows that when $s_{t-1} = L$ and $s_t = L$, only the self probability P_s should be allowed, to account for the fact that the last chord of a structural segment can – and most likely will – last more than one time step. The other probabilities are set to zero.

Since we already established that a key change implies a chord change, we can neglect the influence of the chords c_{t-1}, c_t in the third term $P\left(k_t|s_t, s_{t-1}, k_{t-1}, c_t, c_{t-1}\right)$. We thus end up with $P\left(k_t|s_t, s_{t-1}, k_{t-1}\right)$. Furthermore, we impose the supplemental constraint that a key change can only occur between segments. Mathematically, this can be expressed as $s_{t-1} = O \Rightarrow P\left(k_t|s_t, s_{t-1}, k_{t-1}\right) = \delta_{k_t, k_{t-1}}$ with δ the Kronecker-delta. For the *inter* key transitions $P(k_t|s_t = O, s_{t-1} = L, k_{t-1})$, we reuse the theoretical model from [15], based on Lerdahl's distance [9] between keys.

In Figure 2 one can find a simplified state diagram of our system, in which states are regrouped by the structure variable. Only the transitions from the point of view of the structure variable are drawn in order not to overload the picture, but the constraints on key and chord transitions are indicated next to the arrows. The names of the structure dependent relative chord transition models are also indicated.

2.2.3 *System complexity*

The result of adding the additional constraints is that the complete transition matrix will have a well-defined, sparse structure. The two upper quadrants consist of block diagonal matrices with N_k blocks of side N_c, the lower left quadrant is dense and the lower right quadrant is a diagonal matrix. In comparison to a system that only estimates keys and chords concurrently, the number of states gets doubled by repeating every key-chord state for $s = O$ and $s = L$. On the other hand, because of the sparsity of the transition matrix, the increase in the number of transitions remains limited. More specifically, the number of transitions is $(N_k N_c)^2 + 2N_k N_c^2 + N_k N_c + 1$, which corresponds in our configuration to an increase of 8% instead of the theoretically maximum of 400% that would be reached for a dense transition matrix. This sparsity is exploited in the implementation of the Viterbi algorithm to limit the increase in computation time.

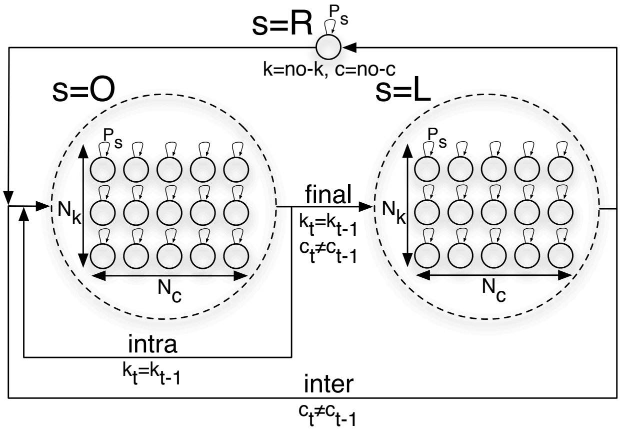

Figure 2. State diagram of the system

2.2.4 *Chroma emission probabilities*

As our chroma features, we use the implementation by Ni et al. [13] known as Loudness Based Chromagrams. These are 24-dimensional vectors that represent the loudness of each of the 12 pitch classes in both the treble and the bass spectrum. They are calculated with a hop size of 23 ms and are afterwards averaged over the interbeat interval. We make the assumption that keys and chords can be independently tested for compliance with an observation and that the structure position is conditionally independent of the observations, such that $P\left(\mathbf{y_t}|s_t, k_t, c_t\right) = P\left(\mathbf{y}_t|c_t\right) P\left(\mathbf{y}_t|k_t\right)$. The chord acoustic probability $P\left(\mathbf{y}_t|c_t\right)$ is modelled as a multi-variate Gaussian with full covariance matrix. Its parameters are trained on the aforementioned "Isophonics" data set. The key acoustic probability $P\left(\mathbf{y}_t|k_t\right)$ is calculated by taking the cosine similarity between the observation vector y_t and Temperley's key templates [18]. These represent the stability of each of the 12 pitch classes relative to a given key.

2.3 The novelty-based component

The other major component of our system is a novelty-based structural segmentation algorithm. We'll use a simple implementation that is conceptually very close to Foote's original proposal [3]. First, a self-similarity ma-

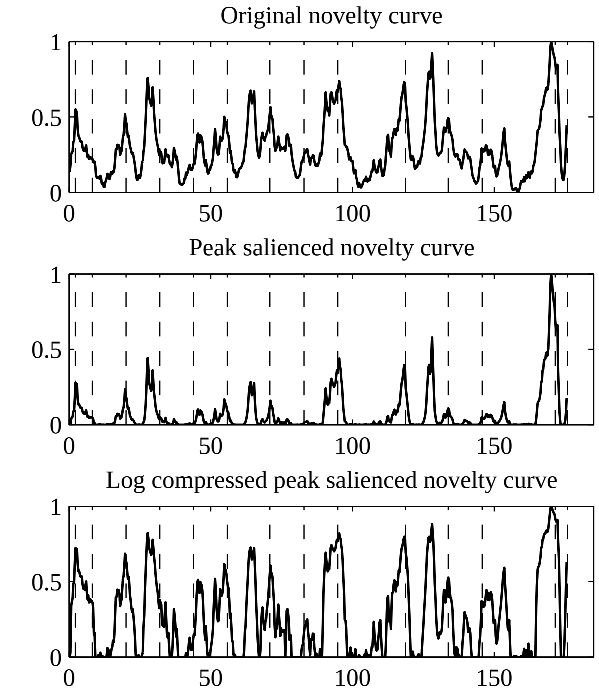

Figure 3. An example of the transformations of the novelty curve

trix is calculated from a sequence of MFCC's and the first four spectral moments (spectral centroid, spread, skewness and kurtosis). The similarity measure that is used is the cosine similarity. From this matrix, a time varying novelty curve is derived by convolving the matrix with a two-dimensional novelty kernel along its diagonal. We use a kernel size of 22.5 s and a step size of 250 ms. In a stand-alone system, the peaks of the resulting novelty curve are detected and structural boundaries are inserted at those positions. In our case however, we'll use the complete curve to calculate the novelty observation probability $P(\mathbf{z_t}|s_t, k_t, c_t)$.

We assume that the novelty observations are conditionally independent of chord and key state, such that we end up with $P(\mathbf{z_t}|s_t, k_t, c_t) = P(\mathbf{z_t}|s_t)$. We thus need to model the novelty observations for each of the possible values of s_t, i.e. O, L or R. By definition, there is a clear relation between the peaks of the novelty curve and a high probability of being in a state that will insert a structural boundary upon leaving it ($s = O$ or $s = R$). Therefore we model the structure acoustic probability for the L and R states as a (half) Gaussian centered on 1. Likewise, the probability of the $s = O$ states will be modelled by a half Gaussian centered on 0. However, we won't use the values of the novelty curve directly as the observations $\mathbf{z_t}$ for the novelty observation probability $P(\mathbf{z_t}|s_t)$. A first remark is that the novelty curves are designed to be used in conjunction with peak picking. Therefore their relative value with respect to surrounding valleys is what matters, and not their absolute value as is desired for our probabilistic system. In order to adapt the novelty values to our use, we therefore first transform the values of the curve. If we represent the values of the novelty curve by $\mathbf{v}$, then the transformation is the following:

$$v'_j = (v_j - \min(v_{j-w} : v_j))\,(v - min(v_j : v_{j+w}))$$

where j is the index of the novelty curve, which has a fixed sample rate of 4 Hz, and w the size of a window around the j-th value. An example of this transformation can be seen in Figure 3, where the original signal is represented in the first row and the processed one in the second. The effect is that valleys now reach all the way down to 0 and that highs indicate peaks with a high salience. As evident in our example, there can be quite a large difference between the saliences of the peaks corresponding to annotated segments. In order to diminish the differences, we apply a log compression, which is subsequently rescaled to the interval $[0, 1]$ for convenience:

$$v''_j = log_{10}\left(1 + \alpha v'_j\right)$$

The result on our example curve is shown in the third row of Figure 3. Finally, we want to account for small deviations of the peak positions with respect to the actual structural boundaries. Therefore we look for extrema of the log-compressed novelty curve in the current beat segment and the segments that are adjacent on each side. For $P(z_t|s_t = L)$ and $P(z_t|s_t = R)$ this will be the maximum and for $P(z_t|s_t = O)$ the minimum, so that the final dimension of $\mathbf{z_t}$ will actually be 2-dimensional: one dimension with the local minima of v'' and one with its local maxima. This step also changes the time scale from a fixed step size of 250 ms to beat-synchronous observations (index j to t).

3. EXPERIMENTAL RESULTS

In this section, we will evaluate our system for various configurations. We evaluate the structural segmentation by calculating the precision $\mathcal{P}(tol)$ and recall $\mathcal{R}(tol)$ between the generated and the annotated structure. They are a function of a tolerance interval tol whose purpose is to allow for small deviations from the desired result to be still considered correct. The precision is defined as the number of estimated boundaries for which an annotated boundary lies within the tolerance interval centered around its position divided by the total number of estimated boundaries. The recall on the other hand, is the relative number of annotated boundaries that have an estimated boundary within its tolerance interval. Both measures are combined in an F-measure $\mathcal{F}(tol)$. These measures are calculated for every song of the data set and are afterwards averaged to give one global result.

We calculate the results for two tolerance intervals, 0.5 s and 3 s, in accordance with the MIREX structure segmentation competition. Both the harmony-based and the novelty-based system were tested separately, as well as the combination of both approaches. For the harmony-based and the combined system, we performed our experiments twice: once with the structure dependent relative chord models derived from the Isophonics data set and once with

	$\mathcal{F}(3s)$	$\mathcal{F}(0.5s)$
harmony-based (Isophonics)	54.72	34.90
harmony-based (Quaero)	52.44	29.47
novelty-based	61.84	33.53
combined (Isophonics)	64.38	35.41
combined (Quaero)	64.13	34.08

Table 2. Results on the Isophonics data set

those from the Quaero set. The results on the Isophonics data can be found in Table 2. We can see that the combination of both approaches has a synergetic effect. Separately, they each have different strengths. The harmony-based approach is better in precisely locating the structure boundaries, as apparent from the $\mathcal{F}(0.5s)$ results, while the novelty-based approach performs better when a larger deviation is allowed. As could be expected, the harmony-based and combined systems work better with the Isophonics relative chord models, since they are perfectly matched with the test set. However, most of the synergy remains when using the models learned on the Quaero set, showing the generality of these models. Additionally, the combined system is also less sensitive to the choice of relative chord models than the harmony-based method is.

4. CONCLUSION

In this paper, we proposed a method for structure estimation by combining 2 different approaches. The first is a traditional way of segmenting structure based on a timbral novelty measure. The second is based on a harmonic analysis that is performed concurrently with the structure estimation. It makes use of chroma features. Together they form a probabilistic system for the simultaneous estimation of keys, chords and structure boundaries. We've shown that the combination of both approaches works better than each of the two systems on its own.

In the future, we will experiment with a post-processing step to extend our resulting structural segmentation into a full structure estimation that includes the identification and labelling of repeated segments. After all, for each of our estimated segments we have a harmonic analysis available that could be used as a feature for the clustering of similar segments, in addition to the more low-level features that are currently used for this task.

5. ACKNOWLEDGEMENTS

This work was partly supported by the Quaero Program funded by Oseo French agency.

6. REFERENCES

[1] M. J. Bruderer, M. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In *Proc. ISMIR*, 2006.

[2] J. Foote. Visualizing music and audio using self-similarity. In *Proc. ACM Multimedia*, 1999.

[3] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. ICME*, 2000.

[4] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. Audio, Speech, Language Process.*, 15(5), 2006.

[5] K. Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal Advances in Signal Processing*, 073205, 2007.

[6] F. Kaiser and G. Peeters. Multiple hypotheses at multiple scales for audio novelty computation within music. In *Proc. ICASSP*, 2013.

[7] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, 1995.

[8] K. Lee. *A system for acoustic chord transcription and key extraction from audio using hidden Markov models trained on synthesized audio*. PhD thesis, Stanford University, 2008.

[9] F. Lerdahl. *Tonal pitch space*. Oxford University Press, New York, 2001.

[10] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2), 2008.

[11] N. C. Maddage. Automatic structure detection for popular music. *IEEE MultiMedia*, 13(1), 2006.

[12] M. Mauch and S. Dixon. Using musical structure to enhance automatic chord transcription. In *Proc. ISMIR*, 2009.

[13] Y. Ni, M. McVicar, R. Santos-Rodríguez, and T. De Bie. An end-to-end machine learning system for harmonic analysis of music. *IEEE Transactions on Audio, Speech and Language Processing*, 20(6), 2012.

[14] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *Proc. ISMIR*, 2010.

[15] J. Pauwels, J.-P. Martens, and M. Leman. Improving the key extraction accuracy of a simultaneous key and chord estimation system. In *Proc. ICME*, 2011.

[16] J. Pauwels and G. Peeters. Segmenting music through the joint estimation of keys, chords and structural boundaries. In *Proc. ACM Multimedia*, 2013.

[17] G. Peeters and H. Papadopoulos. Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6), 2011.

[18] D. Temperley. *The cognition of basic musical structures*. MIT Press, 1999.